Yuvraj Singh Bhadoria

Gurgaon, IN | LinkedIn | +91-6264569114 | yuvrajsinghbhado2030@gmail.com| Github

SUMMARY

AI Engineer specializing in **LLM-powered systems and production-grade RAG pipelines** with **3.5+ years** of **experience building scalable AI/ML solutions**. Expertise in **transformer models, hybrid retrieval, prompt optimization, and evaluation pipelines**, delivering **low-latency, cost-efficient, and reliable AI system**s. Skilled in **cloud deployment (AWS, Azure), vector databases**, and **end-to-end model lifecycle management in remote-first environments**.

SKILLS & INTERESTS

AI & Machine Learning: LLM (GPT, Llama, Claude), NLP, Transformers, RAG, Vector Embeddings, Fine-tuning, Prompt Engineering Deep learning, Model Deployment, A/B Testing, Model Evaluation and Optimization

ML Frameworks & Libraries: PyTorch, Scikit-learn, Hugging Face, Langchain , Keras, Xgboost, Langraph

Programming & Tools: Python, SQL, FastAPI, Streamlit, Git, Docker, Jupyter Notebooks ,Pandas,MLOps

Data Science Techniques: Feature Engineering, EDA, Statistical Analysis, Time Series Analysis, Data Preprocessing

Databases & Cloud: Azure (Cognitive Search), AWS (SageMaker, Bedrock), Vector Databases (Pinecone,Chroma,FAISS,Qdrant) ,Big Query

WORK EXPERIENCE

Bank of America — **Gurgaon , IN**

AI/ML Engineer — *June 2025-Present*

Skills – *SQL*, Python, LangChain, RAG, Prompt Engineering, Microsoft Copilot API, NLP, PyTorch, FastAPI, Docker, Azure, Vector Databases, Git ,Ml

- Developed and deployed **Ericca**, a production-grade AI assistant for IT and HR automation, **reducing support workload by 40%** and response time from **24 hours to 10 minutes**.
- Built a **Retrieval-Augmented Generation (RAG) system** over **15,000+ documents** using LangChain and vector databases, increasing retrieval accuracy from **52% → 88%**.
- Defined and implemented **prompt engineering and evaluation frameworks** for LLMs, improving response quality and reducing API token usage by **25%**.
- Designed secure, scalable **data pipelines** for preprocessing, PII masking, and production deployment of AI workflows using **Python, Pandas, FastAPI, Docker, and cloud infrastructure**.

Policy bazaar — **Gurgaon , IN**

***Senior* Data Scientist** — *March 2025-May 2025*

Skills – Python, SQL, PySpark, AWS, MySQL, ETL Pipelines, Database Optimization , Feature Engineering

- Optimized **Python-SQL ETL pipelines** for ML models and business analytics, reducing end-to-end processing time by **40%** and enabling faster experimentation and insights**.**
- Designed **database architecture and queries**, improving execution speed by **95% (1 hr → 3 min)** and increasing server efficiency by **30%**.

Infosys — **Gurgaon, IN**

Data Scientist — *Dec 2022-Feb 2025*

Skills – Python, SQL, BigQuery, Statistical Modeling, Power BI

- **Led predictive modeling and feature engineering** on large-scale energy datasets (1M+ records), improving **forecast accuracy by 15%** and enabling **data-driven business and operational decisions.**
- **Converted complex datasets into actionable insights**, building **30+ dashboards and analytics reports** that supported **600+ requirements**, drove **25% operational growth**, and ensured **reproducible, scalable models.**

PROJECTS

Enterprise Document Q&A System with Multi-Agent Architecture — **Project Link**

Technologies – Python, FastAPI, Docker, FAISS, BM25, Groq (Llama-3),Cross encoder , Sentence-Transformers

- **Overview:** Built an **enterprise-grade document search and RAG system** using **hybrid retrieval** (semantic + lexical) with **cross-encoder reranking** and **safety guardrails** including confidence-gating and automated groundedness validation.
- **Outcome:** Achieved **98%+ retrieval precision** and **near-zero hallucination** on WikiQA/GovReport datasets, while reducing irrelevant responses by **40%** using **threshold-based refusal logic**.

Customer Churn Prediction System Using Machine Learning — **Project Link**

Technologies: Python, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn, Logistic Regression, Random Forest, SVM, KNN, Decision Trees

- **Overview:** Created a banking customer **churn prediction system** using **5 supervised ML models**, **performing EDA, feature engineering, label encoding, and feature scaling**, with comparative **evaluation via confusion matrices.**
- **Outcome:** Achieved **86% accuracy with Logistic Regression, Random Forest, and Decision Tree models** (KNN: 78%, SVC: 79%), and built a **CLI-based prediction** interface with joblib for real-time churn forecasting.

LLM Fine-Tuning with Direct Preference Optimization (DPO) — **Project Link**

Technologies: Python, PyTorch, HuggingFace Transformers, FSDP, Anthropic-HH Dataset, Weights & Biases ,Mixed Precision Training

- **Overview:** Implemented an LLM alignment pipeline using **Direct Preference Optimization (DPO)** to **fine-tune Pythia models (2.8B–6.9B parameters)** on human preference data, leveraging **FSDP and bfloat16 mixed precision** for efficient multi-GPU distributed training.
- **Outcome: Accelerated model training by 50%** and improved model alignment through **hyperparameter optimization (beta: 0.1-0.5)** and **preference-based fine-tuning** on Anthropic-HH dataset.

EDUCATION

Rustamji institute of Technology — **Gwalior, IN**

BTech (Automobile Engineering) — *Aug 2018-May 2022*

CGPA *-8.57*